SPRINGONE2GX
WASHINGTON, DC

Who Needs Batch?

By Michael Minella and Gunnar Hillert
@michaelminella, @ghillert




Gunnar Hillert

Twitter: @ghillert
Website: http://blog.hillert.com
Conference: http://devnexus.com

Michael Minella

Twitter: @michaelminella
Podcast: http://javaOffHeap.com
or @OffHeap
Website: http://spring.io

https://github.com/ghillert/s1-2gx2015-who-needs-batch

WE HAVE AN ANNOUNCEMENT

WE HAVE A DISEASE

springone 2GX

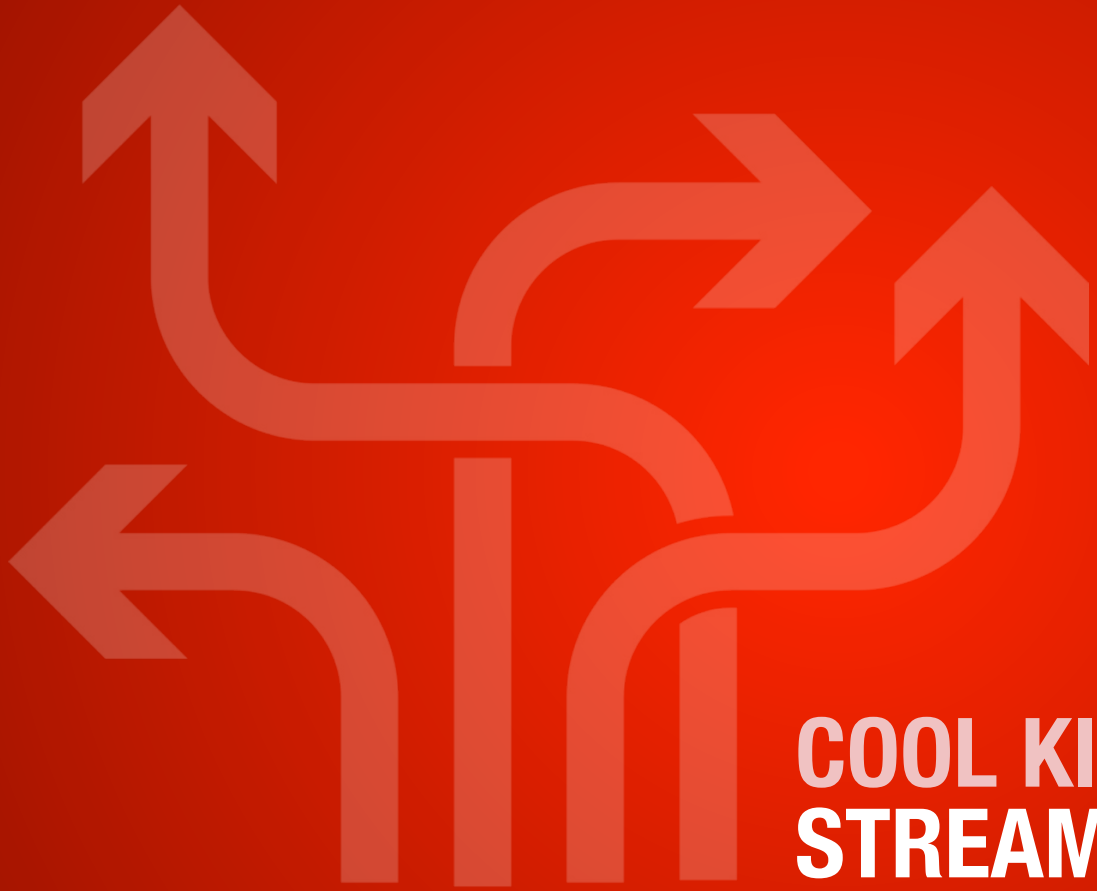
OBJECTIVIUS SHINIUM SYNDROMUS


SHINY OBJECT
SYNDROME

IT MAKES US
INNOVATE

ROAD TO RECOVERY

springone 2GX



COOL KIDS ARE DOING STREAM PROCESSING


Flink


Stream Analytics


Google Cloud Dataflow


Tigon


PULSAR


sqlstream


STORM

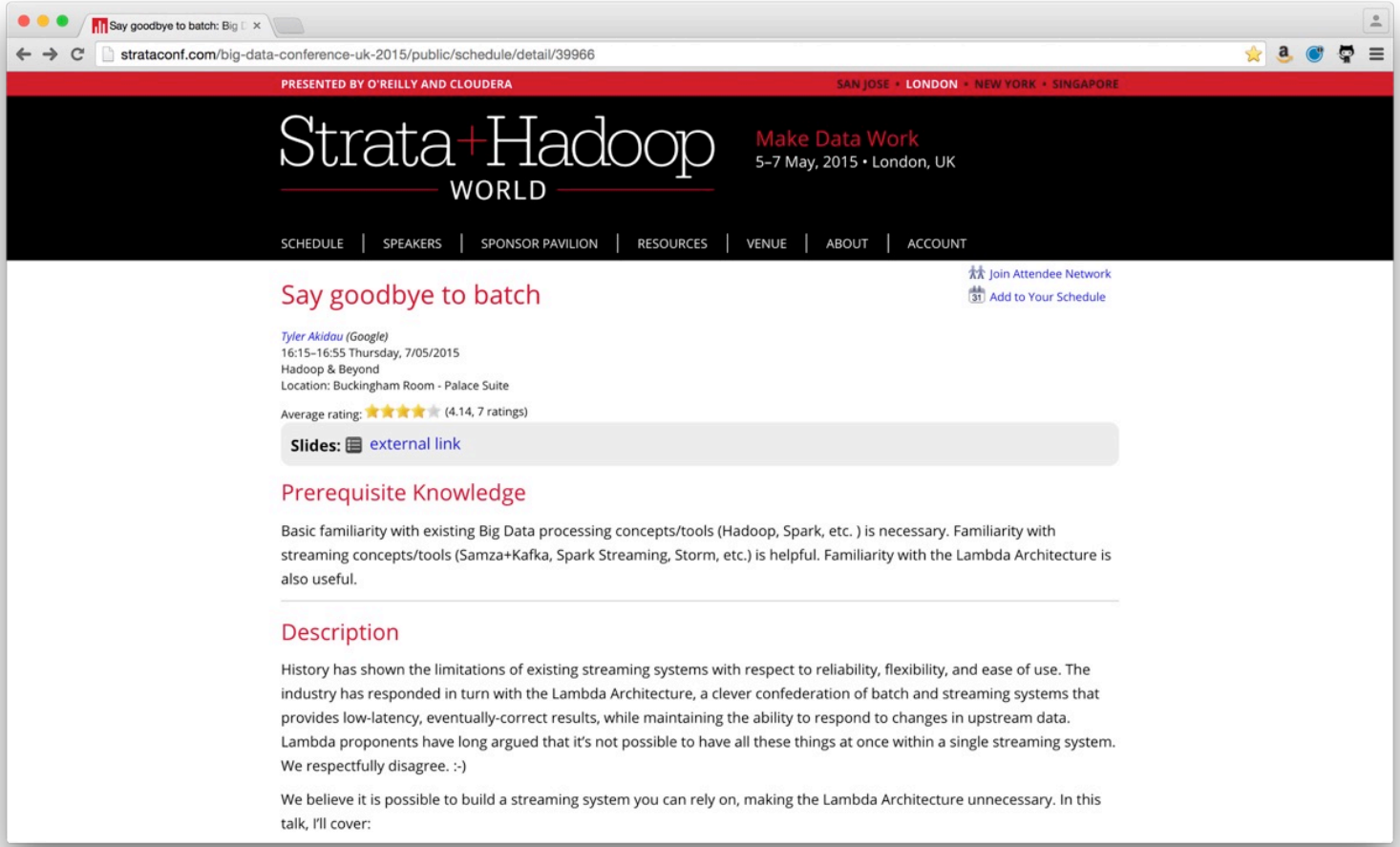

S4
distributed stream
computing platform

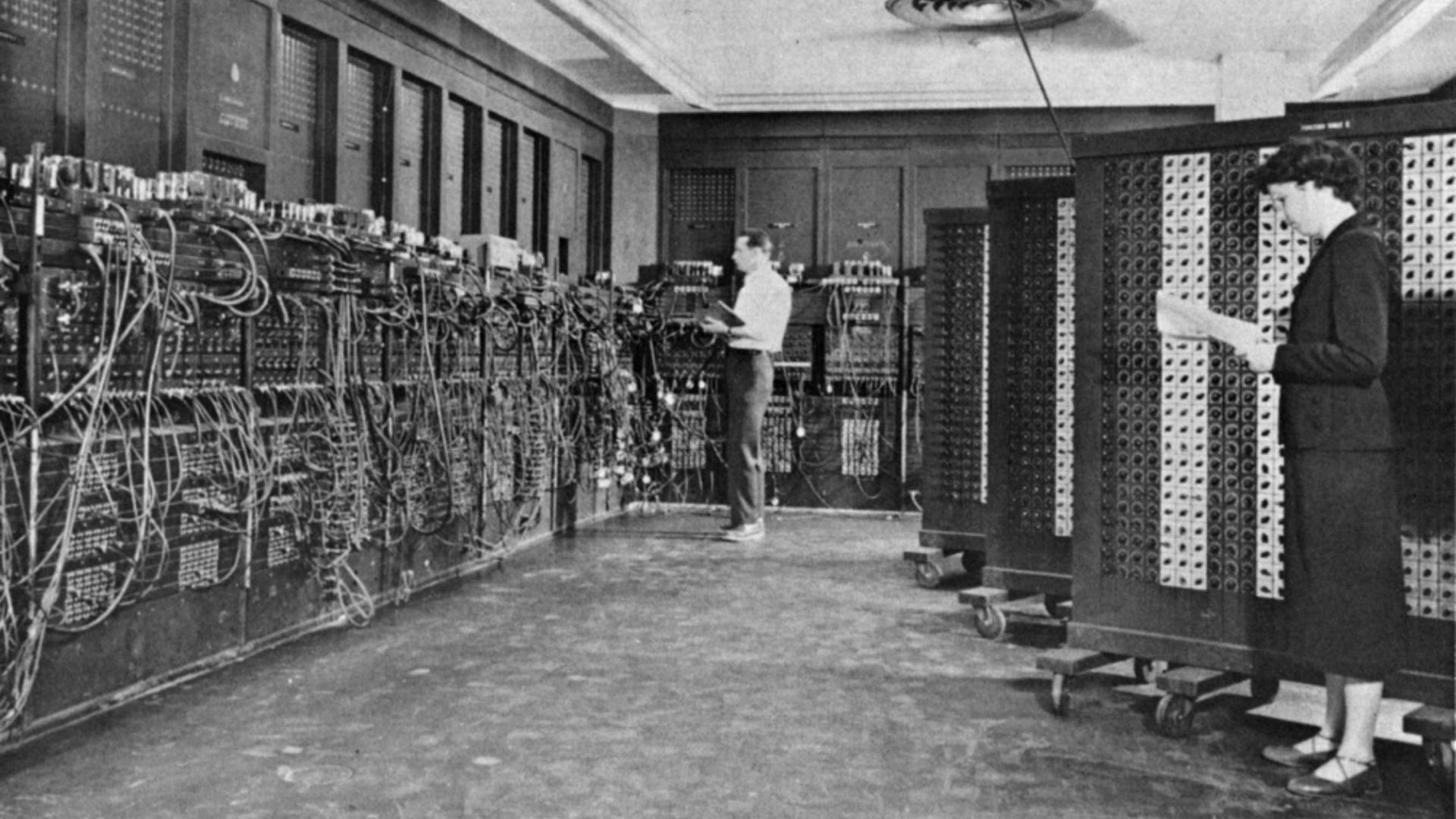

Spark
Streaming


samza


EsperTech

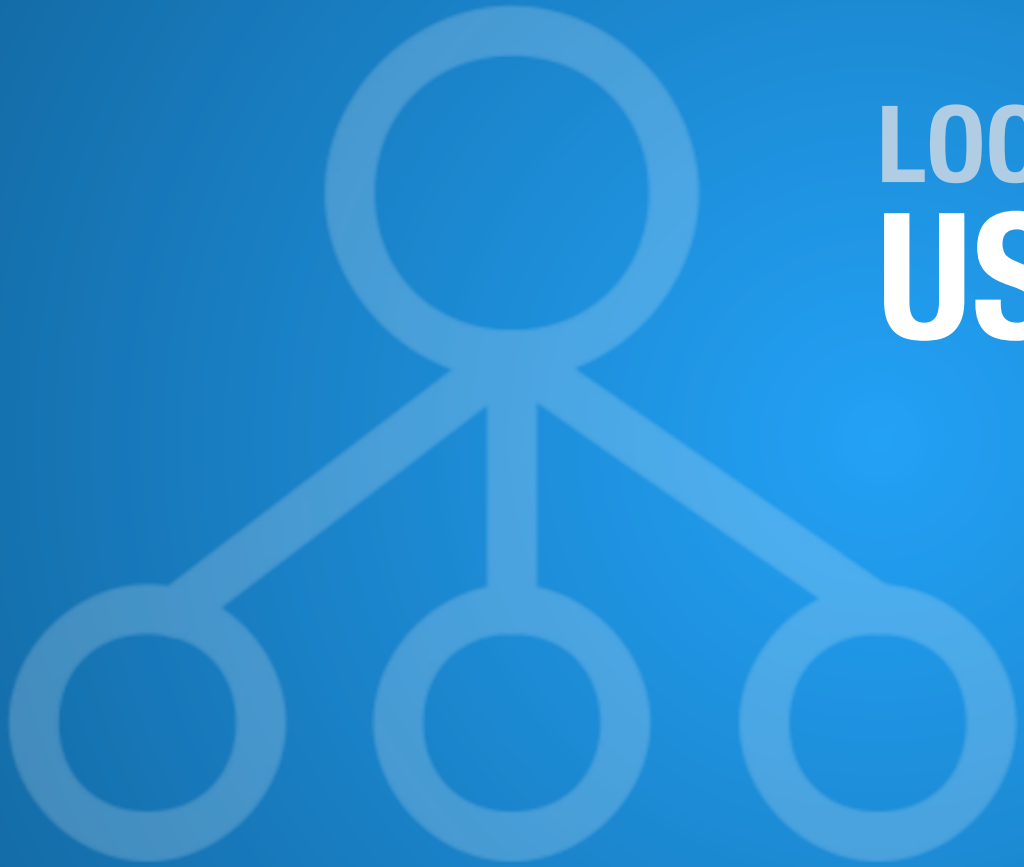

springone 2GX





LET'S TAKE A CLOSER LOOK

BATCH AND STREAM DEFINITIONS



LOOK AT COMMON USE CASES

CONCLUSIONS

DEFINITION:
STREAM PROCESSING

A system for processing an unbounded set of data in an asynchronous manor.

DEFINITION:
BATCH PROCESSING

DEPENDS ON
WHO YOU ASK

“Batch … applications run … as special cases of stream processing applications”

- Apache Flink Documentation

Batch processing is the processing of a **bounded data set** without **interruption** or **interaction**

KEY DIFFERENCES

1 BOUNDED VS UNBOUNDED

2 SYNCHRONOUS VS ASYNCHRONOUS

3 STATEFUL VS STATELESS

STATE OF STREAMING


Flink


Stream Analytics


Google Cloud Dataflow


Tigon


PULSAR


sqlstream


STORM


S4
distributed stream
computing platform


Spark
Streaming


samza


EsperTech


springone 2GX


Flink

Stream Analytics


Google Cloud Dataflow

Tigon

PULSAR


sqlstream

STORM


S4
distributed stream
computing platform


Spark
Streaming

samza


EsperTech


springone 2GX

STATE OF STREAMING

EasyBatch

JSR-352

WHY THE
END OF BATCH?

LATENCY

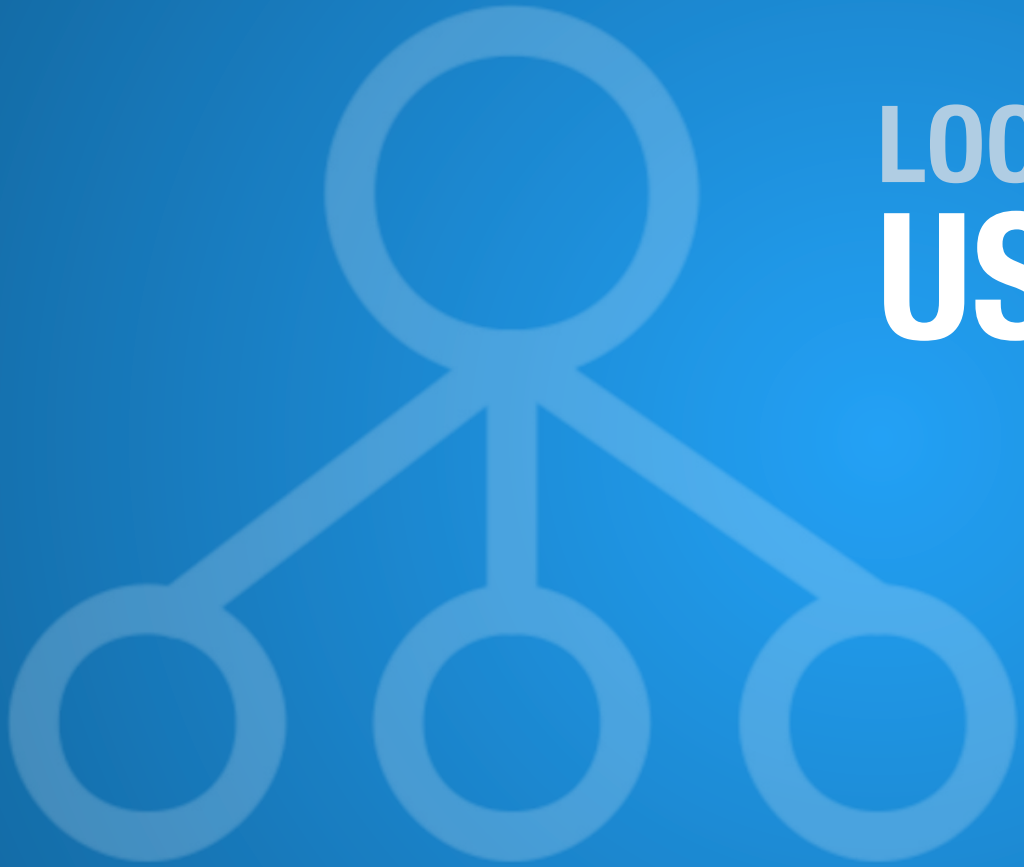
WHAT DOES THE END OF BATCH MEAN?

LIMIT LATENCY

“ENTERPRISE GRADE”

EasyBatch

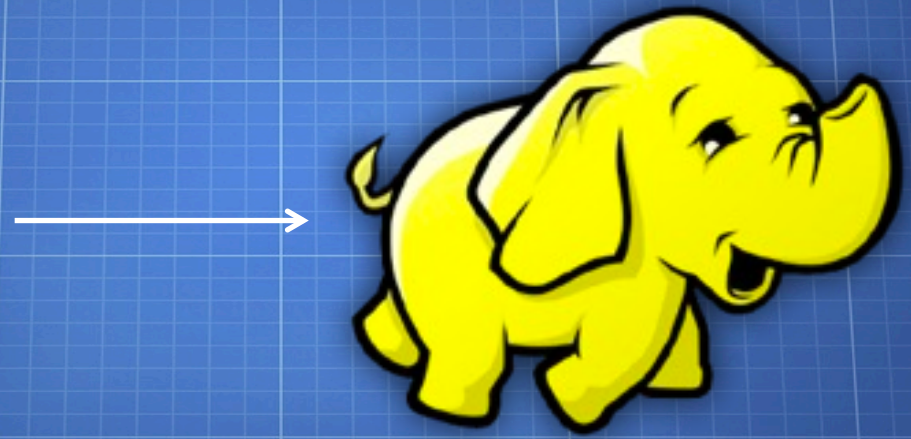
JSR-352



LOOK AT COMMON USE CASES

E.T.L.

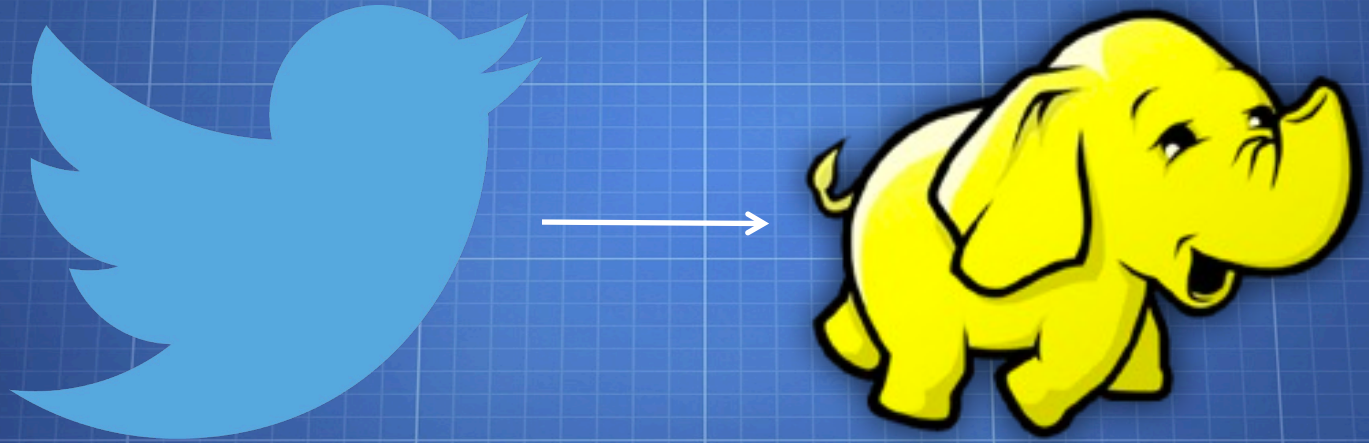
DATABASE TO HDFS



DEMO

STREAMING CAN BE USEFUL IN INGESTION USE CASES

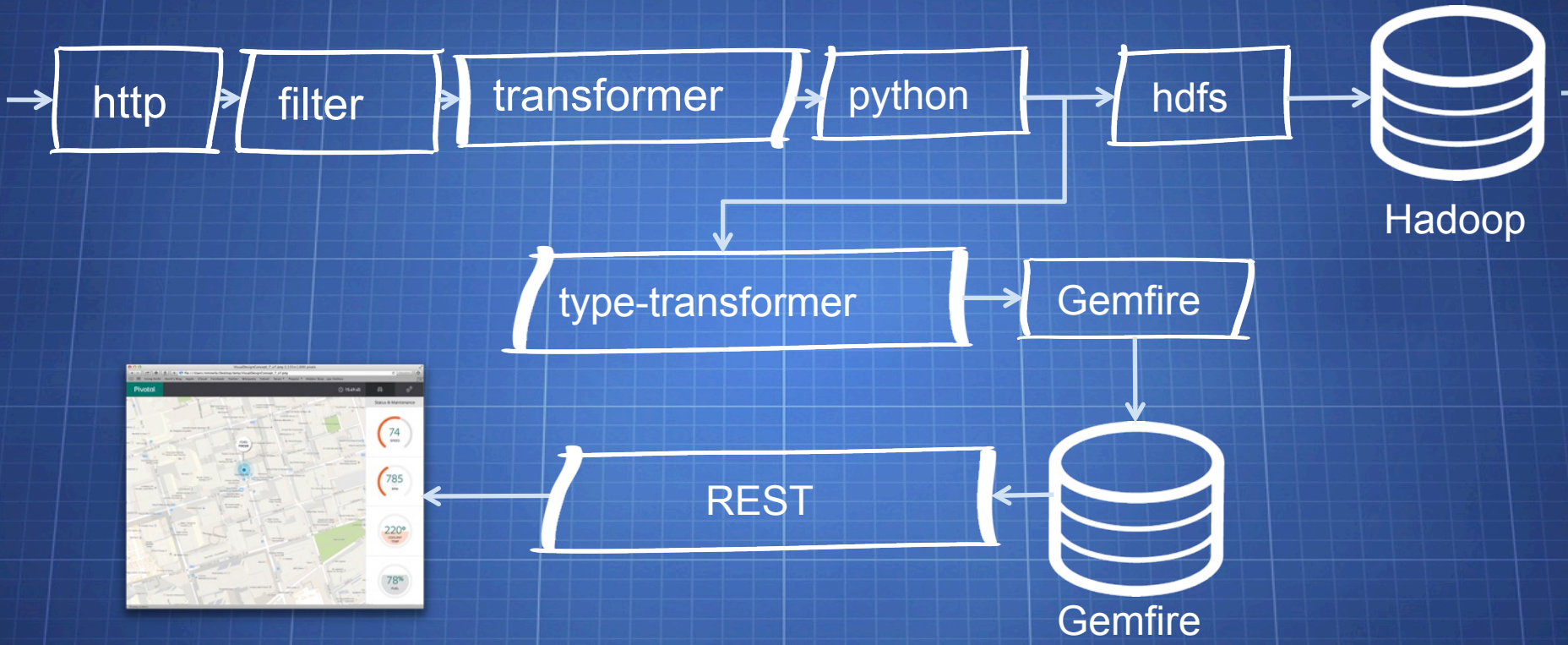
TWITTER TO HDFS



DATA SCIENCE

BATCH TRAINING
PREDICTIVE MODELS

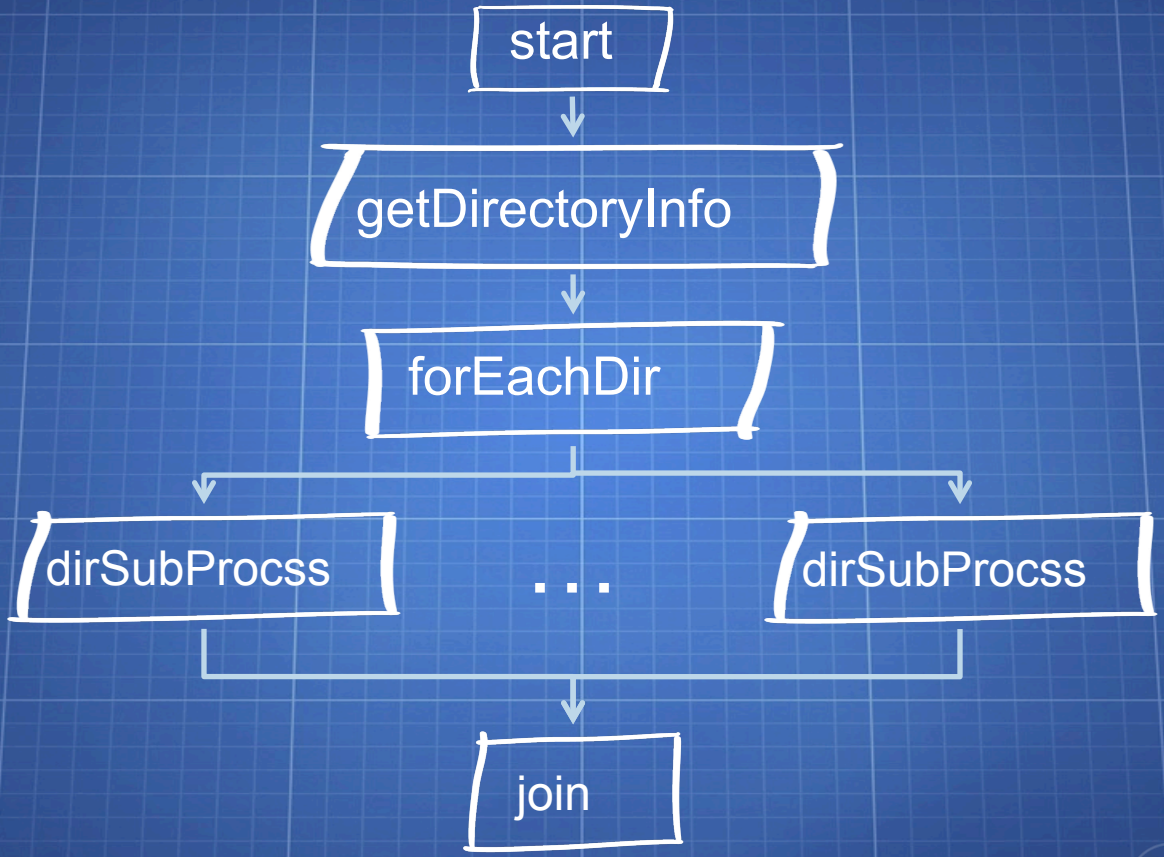
CONNECTED CAR

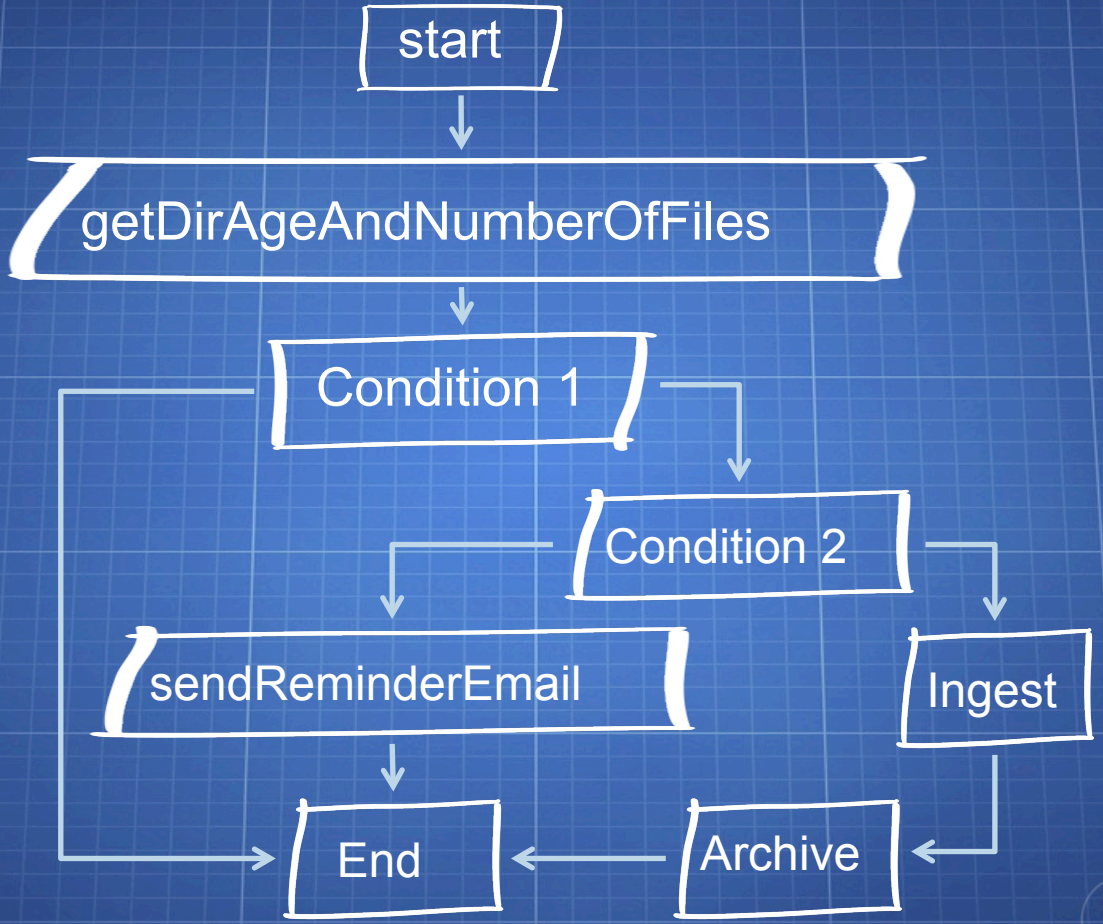

http
filter
transformer
python
hdfs
Hadoop
type-transformer
Gemfire
REST
Gemfire

STREAMING APROXIMIATIONS EXIST

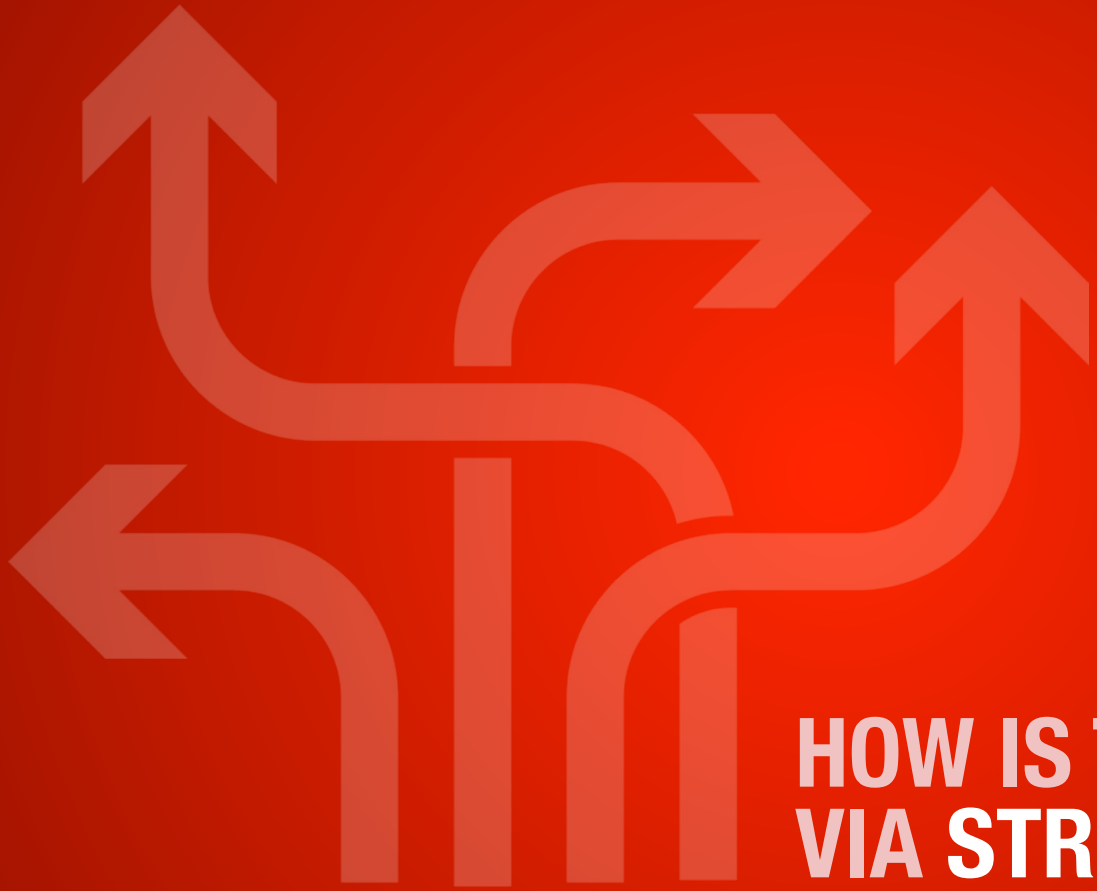
BREADTH AND ACCURACY DON'T MATCH BATCH

WORKFLOW ORCHESTRATION





DEMO



HOW IS THIS DONE VIA STREAMING?

NON-INTERACTIVE PROCESSING

SAME AS ETL PROCESSING

PORT SCANNER

DEMO

STREAMING DOESN'T MAKE SENSE

STREAMING
DOES MAKE SENSE

WHERE LATENCY IS A PRIORITY

DATA LOSS
IS ACCEPTABLE

UNBOUNDED
DATASET

BUT DOES NOT
IN OTHER USE CASES

HIGHLY COMPLEX

ERROR HANDLING
NOT AS ROBUST

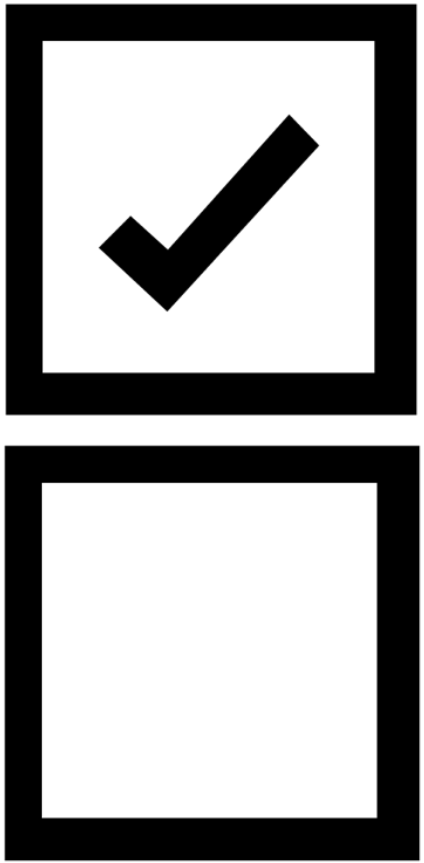
DATA GUARANTEES NOT AS ROBUST

WHERE BATCH IS BETTER

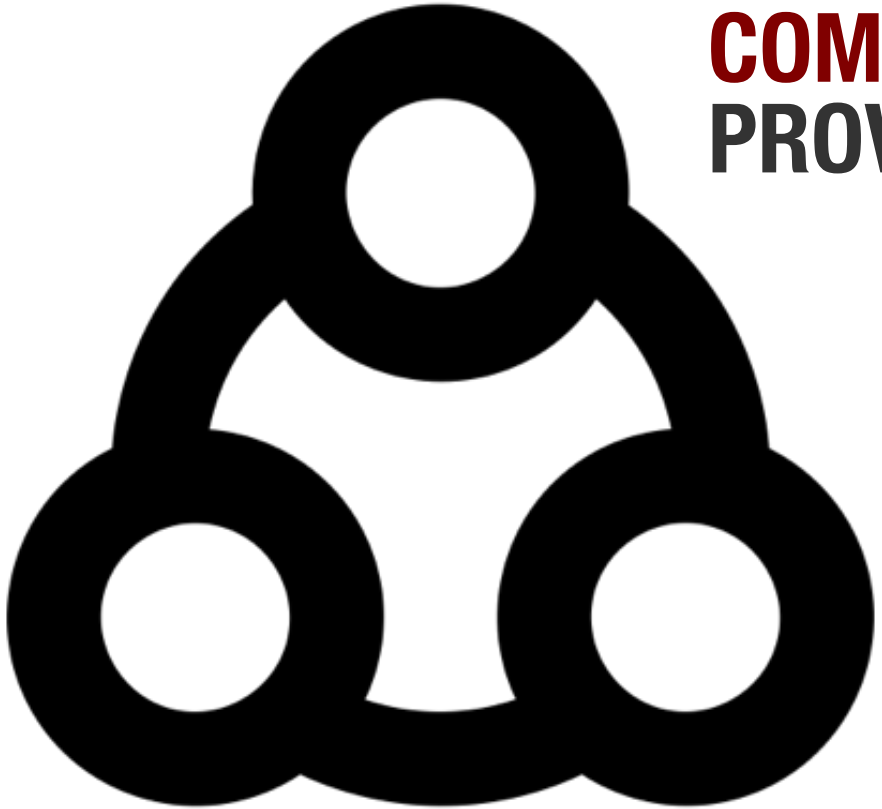
FINITE DATASET

ROBUST ERROR HANDLING

RESOURCES ARE A CONCERN



NOT EITHER OR

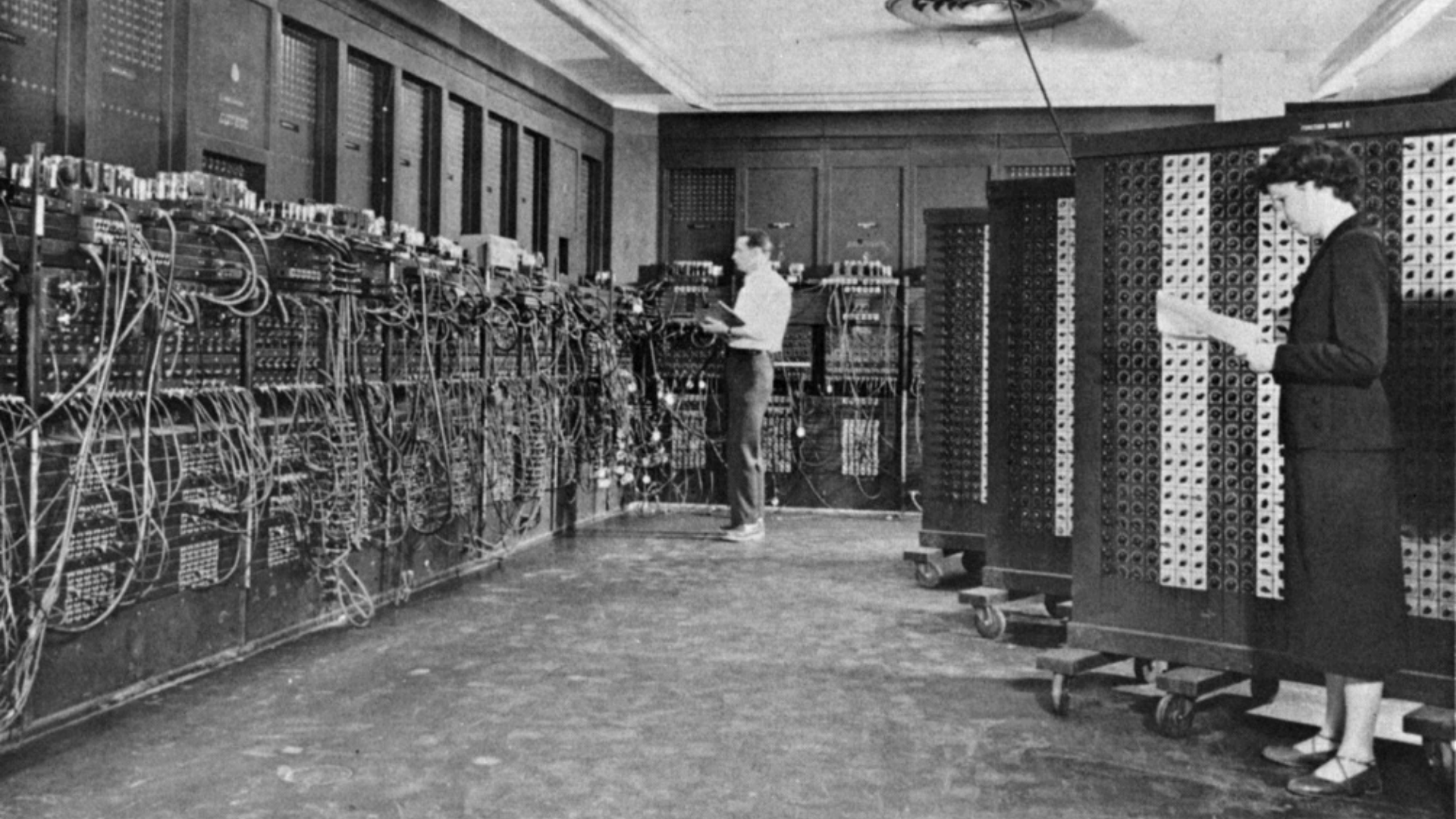


COMBINING THE TWO PROVIDES THE MOST POWER

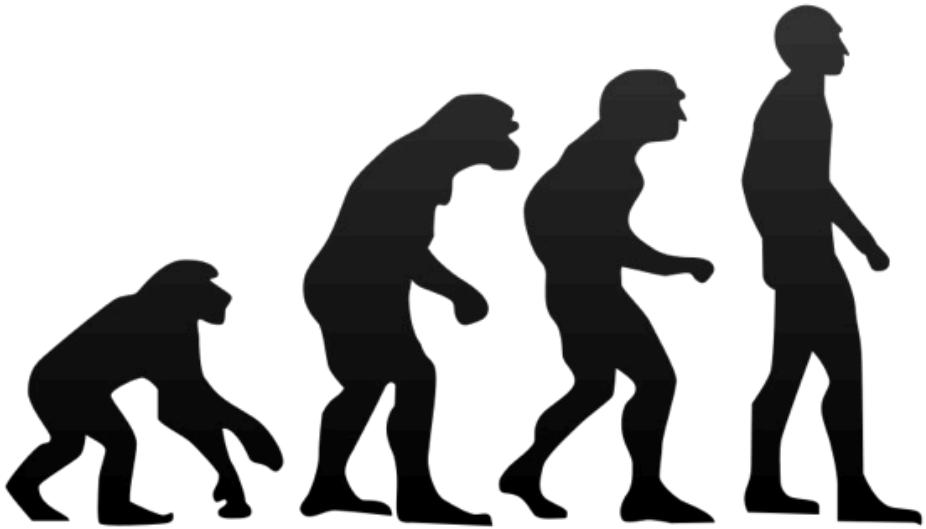
E.T.L.v2

Database to HDFS v2

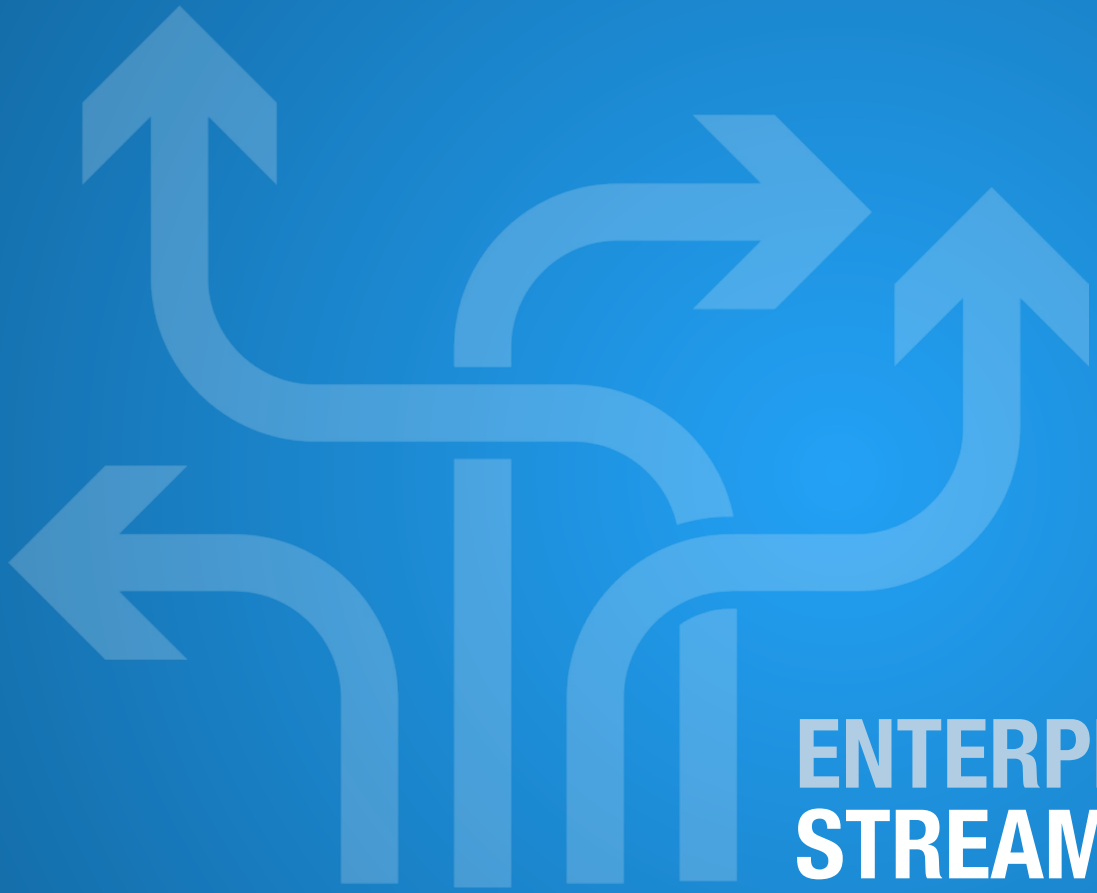
DEMO



BATCH HAS LASTED BECAUSE IT MAKSE SENSE



SPRING XD/DATA FLOW
BEST OF BOTH WORLDS



ENTERPRISE GRADE STREAM PROCESSING

BEST BATCH OPTION
ON THE JVM

SPRINGONE2GX
WASHINGTON, DC

Learn More. Stay Connected.

 @springcentral

 Spring.io/video

Check out Spring Cloud Data Flow on https://spring.io

Apache Spark for Big Data Processing – Salon N-P Up next!
High Performance Stream Processing – Salon N-P 8:30AM Tomorrow
Spring Integration Extensions Ecosystem – Here! 12:45 Tomorrow